Multiphase Inverse Modeling
Lecture Notes

Stefan Finsterle

Lawrence Berkeley National Laboratory
Earth Sciences Division
University of California
Berkeley, CA 94720

Phone: (510) 486-5205
Fax: (510) 486-5686
E-mail: SAFinsterle@lbl.gov

April 20, 2000

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Scope

Predicting multiphase flow and transport processes in the subsurface by means of numerical simulation involves the following steps:

1. Developing a conceptual model of the natural system;

2. Assigning values to the input parameters;

3. Running the model, i.e., predicting the system state;

4. Interpreting the results and assessing the uncertainty of the predictions.

The first step is the most difficult and also most important task, because the conceptual model provides the basis for all the subsequent steps. Errors in the conceptual model usually have the largest impact on model predictions.

In multiphase flow modeling, the second step (assigning parameter values) can be tedious because of the large number of parameters that enter the model. Moreover, the physical interpretation of these parameters is often ambiguous and they are difficult or even impossible to measure directly. Parameters can be estimated by automatically calibrating the model against measured data. Inferring model-related parameters from observations is termed *inverse modeling.* Inverse modeling deals with parameters and their sensitivities. Consequently, information regarding the significance of the parameters and their impact on model predictions is obtained as a by-product of inverse modeling.

This lecture is intended to introduce inverse modeling concepts for applications in multiphase flow and transport simulations. While inverse modeling can be discussed in the jargon of applied mathematics and mathematical statistics, a more practical approach is employed here, taylored to the needs of engineers who are interested in calibrating numerical models against observed data. The following issues will be discussed:

- Fromulation of the forward problem;
- Measurement errors and the stochastic model;
- Maximum likelihood and the objective function;
- Minimization algorithms;
- Residual and error analysis;
- Uncertainty propagation analyses.

An introductory example will be given in Section 1.3, followed by a more detailed discussion of the elements described above. In the following section, inverse modeling is positioned within the overall framework of mathematical modeling.

1.2 Inverse Modeling—The Big Picture

Parameter estimation, history matching, model calibration, and inverse modeling are terms describing essentially the same technique with a slightly different objective in mind. *The ultimate goal of inverse modeling is to assess the best model and its parameters for predicting the behavior of a dynamic flow system.* It is obvious that the reliability of these predictions depends on the appropriateness of the conceptual model and the parameters entering the model. One should keep in mind that it is the intended use of the model that determines the required degree of model sophistication as well as the level of accuracy with which the parameters are to be estimated. In this overall scheme, parameter estimation is only one, albeit important step in the process of model development.

Inverse modeling consists of estimating model parameters from measurements of the system response made at discrete points in space and time. Automatic model calibration can be formulated as an optimization problem that is solved in the presence of uncertainty. Uncertainty is a result of the available observations being incomplete and exhibiting random measurement errors.

The parameters to be estimated consist of selected coefficients in the governing flow equations and may include hydrogeologic and thermophysical properties, initial and boundary conditons, as well as parameterized aspects of the conceptual model. The interpretation of these parameters depends on the model structure and the overall purpose of the model. In this sense, the parameters are strictly to be seen as *model parameters* (or *model-related* parameters) rather than parameters of the geologic formation. Estimating parameter values from measurements therefore relates the real system to its idealized representation.

Inverse modeling involves several iterative steps. Given a conceptual model of the physical system, the results of parameter estimation may indicate that the underlying model structure has to be modified. This process of iteratively updating the conceptual model and its parameters is sometimes referred to as *model identification*. We will focus here on the more narrow aspect of inverse modeling, namely parameter estimation by automatic model calibration.

If automatic model calibration makes use of a gradient-based minimization algorithm, the sensitivity of the calculated system response with respect to the input parameters is evaluated. This information can also be used to study the appropriateness of a proposed experimental design and to analyze the uncertainty of model predictions. A computer program for inverse modeling therefore provides information to support three types of applications:

- Sensitivity analysis;
- Parameter estimation;
- Uncertainty propagation analysis.

All three application modes are of practical significance:

- Sensitivity analyses supply the measures needed to optimize the design of a laboratory experiment or field test. They also help identify the parameters that excert the greatest impact on model predictions. These are the parameters that must be determined with the lowest possible estimation uncertainty.
- Parameter estimation by inverse modeling overcomes the time- and labor-intensive tedium of trial-and-error model calibration. More important, the error analysis provides insight into the uncertainty of the estimated parameters and reveals parameter correlations. Predictability can be improved when relying on effective, model-related parameters estimated by inverse modeling.
- The quality of simulation results can be assessed by propagating the uncertainty of the input parameters through the prediction model.

1.3 Introductory Example

The process of parameter estimation by automatic model calibration is illustrated in the following example, which is described in detail in *Finsterle and Persoff*[1997] [9]. A laboratory experiment was designed to estimate permeability and porosity of very tight rock samples. A schematic of the experimental apparatus is shown in Figure 1.1. A rock sample is dried and placed in a sample holder, which is attached to two gas reservoirs. To conduct a test, the upstream reservoir is rapidly pressurized using nitrogen gas to a value about 300 kPa above the initial pressure of the system. Gas starts to flow through the sample, and the change of pressure with time is observed in both reservoirs.

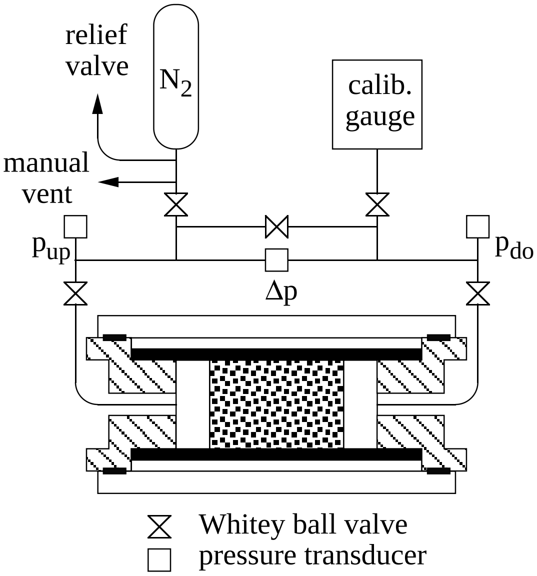

Figure 1.1: Gas-pressure-pulse-decay apparatus

The process of inverse modeling involves the steps listed in Table 1.1. We follow these steps for our specific example:

1. As part of model conceptualization, the relevant physical processes have to be identified, mathematically described, and implemented into a numerical code. In this example, it is sufficient to consider single-phase gas flow. Because of the small porosity of the rock sample, Klinkenberg [13] gas slip flow may become significant and has to be taken into account. The gas flow term is given by:

$$\mathbf{F}_{\rm g} = -k\left(1+\frac{b}{p}\right)\frac{\rho}{\mu}\nabla p \qquad (1.1)$$

where $\mathbf{F}_{\rm g}$ is gas flux [kg s^{-1} m^{-2}], k is absolute permeability [m^2], b is the Klinkenberg factor [Pa], ρ is density [kg m^{-3}], μ is dynamic viscosity [Pa s], and p is pressure [Pa]. Note that density and viscosity of a gas are functions of pressure and temperature as well as relative humidity. This flow equation and the appropriate equation-of-states

Table 1.1: Inverse Modeling Procedure: Major Steps

Step	Description	Issue
1	Develop a numerical model of the experiment.	Model conceptualization
2	Select parameters to be estimated.	Parameterization
3	Select reasonable initial parameter values.	Prior information
4	Select data and identify calibration points in space and time.	Calibration points
5	Assign uncertainties to calibration points.	Stochastic model
6	Calculate system response.	Forward simulation
7	Compare calculated with observed system response.	Objective function
8	Update parameters to decrease difference between observed and calculated system response.	Minimization algorithm
9	Repeat step 6–8 until no further improvement of the fit can be obtained.	Convergence criteria
10	Analyze residuals and estimation uncertainties.	Error analysis

enter the mass and energy balance equations solved by the numerical simulator TOUGH2 [17] [18].

2. The parameters to be estimated are the porosity ϕ, the absolute permeability k, and the Klinkenberg factor b. Since both k and b are expected to vary over many orders of magnitude, the logarithm of these parameters will be estimated. Furthermore, logarithmic transformation makes the inverse problem more linear and prevents the parameters from becoming negative. The three parameters are summarized in a vector $\mathbf{p}$ of length $n = 3$.

3. The initial parameter values are chosen as follows: $\phi = 0.015, \log(k) = -19.0$, and $\log(b) = 7.0$. This example is very insensitive to changes in the initial parameter guesses.

4. Observations available for model calibration are the pressure data in the upstream and downstream reservoirs, respectively. 30 calibration points, logarithmically spaced in time, are selected. The total number of calibration points is therefore $m = 60$.

5. The measurement errors of the pressure data are assumend to be uncorrelated and on the order of $\sigma_{z_i} = \sigma_z = 1000$ Pa, $\quad i = 1, \ldots, m$. The covariance matrix $\mathbf{C}_{zz}$ is a matrix of dimension $m \times m$ with σ_z^2 on the diagonal and zeroes elsewhere.

6. The experiment is simulated using a numerical model. The dash-dotted lines in Figure 1.2 show the pressure transient in the upstream and downstream reservoirs obtained with the initial parameter set.

7. The difference between the model calculation and the data as seen in Figure 1.2 is measured by the objective function. The standard objective function is the sum of the squared weighted residuals:

$$S = \mathbf{r}^T \mathbf{C}_{zz}^{-1} \mathbf{r} = \sum_{i=1}^{m} \frac{r_i^2}{\sigma_i^2} \tag{1.2}$$

   where $\mathbf{r}$ is a vector of length m holding the residuals, i.e., the differences between the observed and calculated pressure at the calibration points (see Equation (1.3) below). Note that weighted least-squares leads to maximum likelihood estimates if the residuals are normally distributed.

8. The minimization algorithm described in Chapter 3 proposes new parameter sets based on the gradient of the objective function with respect to parameter vector $\mathbf{p}$.

9. If a certain convergence criterion is met, go to Step 10, otherwise repeat Step 6 with the updated parameter vector. The best fit obtained after a few iterations is shown in Figure 1.2 (solid lines), matching the observed data (symbols) almost perfectly.

10. The error and residual analysis will be discussed in detail in Chapter 4. In this example, the covariance matrix of the estimated parameters reveals a very high correlation between permeability and Klinkenberg parameter, leading to unacceptable estimation uncertainties despite the perfect match. This demonstrates the importance of the error analysis. The solution to this specific problem of ill-posedness is presented in [9].

The example above and Figure 1.3 illustrate the process and main elements of inverse modeling, which will be discussed in more detail in the following chapters. The example also demonstrates that the parameters of interest may not be identified despite a perfect match. The error analysis suggests that the design of the experiment should be changed in order to reduce the correlation between $\log(k)$ and $\log(b)$. The solution to the ill-posedness of the problem is to simultaneously invert data from multiple

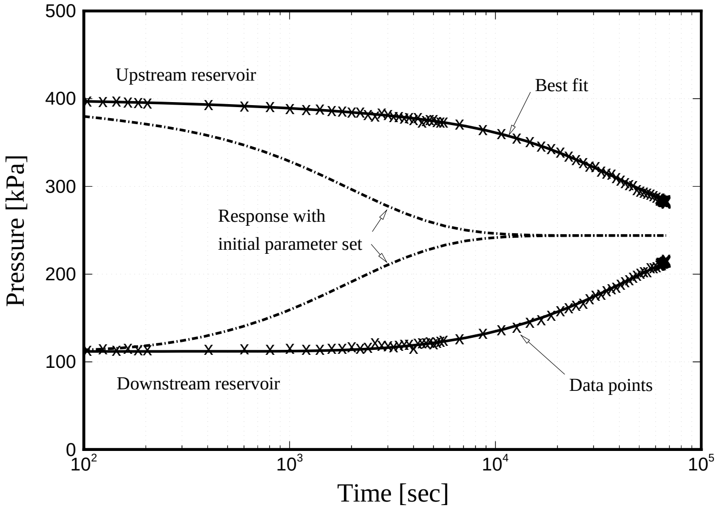

Figure 1.2: Comparison between measured and calculated pressure transient curves with the initial and final parameter set.

gas-pressure-pulse-decay experiments performed on different pressure levels [9].

The concept of inverse modeling as outlined here has been implemented in a computer code named iTOUGH2 [5] [6][7]. iTOUGH2 is based on the TOUGH2 [17][18] numerical simulator for nonisothermal flows of multicomponent, multiphase fluids in porous and fractured media.[1]

1.4 Definitions

1.4.1 Typing conventions

- Scalars are represented by plain characters, e.g., k_{rl}.
- Vectors are lower-case bold characters, e.g., $\mathbf{p}$.
- Matrices are upper-case bold characters, e.g., $\mathbf{J}$.
- Elements of vectors or matrices are scalars with an index, e.g., p_3 or J_{ij}.
- Measured quantities are indicated with an asteriks (*), e.g., the residual $r = (z^* - z)$ is the difference between the measured and the calculated value of variable z; k^* is a measured permeability value (prior information), whereas k is its estimate from inverse modeling.

[1] Information about iTOUGH2 can be obtained from the Web at `http://www-esd.lbl.gov/iTOUGH2`.

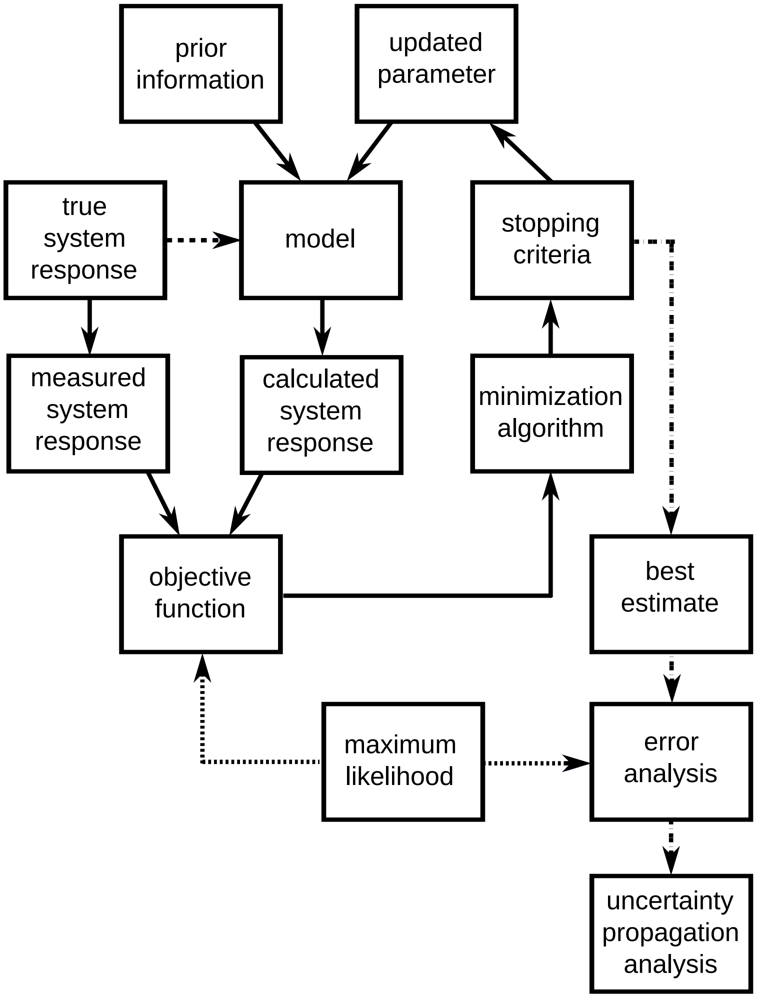

Figure 1.3: Inverse modeling flow chart.

1.4.2 Parameter vector p

The parameter vector **p** of length n contains the parameters to be estimated by inverse modeling. The parameters must be *input* parameters of the simulation model (e.g., TOUGH2). For example, the elements of a parameter vector of length $n = 3$, $\mathbf{p} = [p_1, p_2, p_3]^T$ could be:

- p_1: absolute permeability along the first principle axis of all elements belonging to material domain SAND.

- p_2: second parameter of default capillary pressure function (e.g., representing van Genuchten parameter α.)

- p_3 initial NAPL saturation for all elements of material domain CLAY.

1.4.3 Vector of observations z

Vector **z** of length m contains dependent, observable variables at discrete points in space and time, which are chosen to be calibration points. Elements of **z** may refer to measured quantities (data) or simulation results. Observable variables must be part of the model *output.* Here are a few examples:

- z_1^* is the actually measured or interpolated pressure at a certain point in space and time. z_1 is the corresponding model output, i.e., the calculated pressure in the corresponding element X and at output time T.
- z_2 is the average NAPL saturation in the entire model domain at time T.
- z_3 is the cumulative liquid flow rate to a pumping well, i.e., a sum of calculated fluxes across the corresponding interfaces in the numerical model.

Both the parameter vector **p** and the vector of observations **z** have finite dimensions (n and m, respectively). For a heterogeneous aquifer with continuously varying properties, the dimension of the parameter vector is theoretically infinite. Similarly, the system response is continuous (pressures vary continuously in space and time). Again, the dimension of vector **z** should be infinite. The problem is solved by discretization. The reduction of the number of parameters is called parameterization, i.e., the aquifer system is subdivided into several subregions with presumably constant properties. The reduction of the vector of observations is achieved by picking discrete points in space and time for calibration. Note that discretization is also necessary for numerical reasons, i.e., the continuous partial differential equation describing multiphase flow is solved by a (discrete) finite-difference approximation in space and time.

The vector of observable variables may also contain parameters. For example, if permeability has been measured on cores in the laboratory, this information can be considered as an additional data point, and treated along with the direct observations of the system response. Such measured parameter values are referred to as *prior information.*

1.4.4 Residual vector r

The residual vector **r** contains the differences between the measured and calculated system response with elements

$$r_i = z_i^* - z_i \qquad i = 1, \ldots, m \tag{1.3}$$

For example, r_i is the difference between the measured and calculated pressure at a certain point in space and time. A special type of residuals are the differences between the measured parameters (prior information) and the estimated parameter values. This difference—appropriately weighted—can be used for regularization of the inverse problem, making the solution more stable and well-posed.

1.5 Course Structure

The course follows the diagram shown in Figure 1.3. However, the key element of inverse modeling—the simulation model that solves the forward problem—is not discussed here. The methods described in the following chapters are general and can be applied to basically any type of process simulation. Nevertheless, we discuss inverse modeling in the context of multiphase flow simulation and make occasional reference to the TOUGH2 and iTOUGH2 simulators.

Chapter 2 provides the statistical background and discusses the stochastic model including the choice of the objective function. Minimization algorithms and stopping criteria are introduced in Chapter 3. The residual and uncertainty analyses are discussed in Chapter 4.

The objectives of inverse modeling and optimization are common to many disciplines, including engineering, science, economics, and manufacturing. Unfortunately, textbooks on the subject are often tailored to specific applications with their own terminology. On the other hand, if considered a topic of applied mathematics, inverse modeling is usually presented in a way that makes it difficult to derive useful tools for an engineer who is interested in solving practical inverse problems. This course attempts to bridge the gap between theroretical considerations of existence, uniqueness, and stability of inverse problems, and the practical need to calibrate multiphase flow models.

Introducions to optimization can be found in a number of textbooks [1][11][15]. Review articles [14][19] in the field of water resources may serve as a starting point for further reading. We specifically mention the series of articles by *Carrera and Neuman* [1986] [2][3][4], which describe the concepts of inverse modeling in a concise manner.

Chapter 2

The Stochastic Model

2.1 Systematic and Random Errors

Recall that inverse modeling consists of estimating parameters by minimizing some norm of the residuals (for example, Equation (1.2)). Residuals as previously defined in Equation (1.3) contain contributions from both measurement and modeling errors. Consider a dataset that is drawn from a true but unknown system response (Figure 2.1). The individual measurement error is defined as the difference between the measured and the true value. The modeling error is defined as the difference between the true and the calculated system response. Since the true system response is unknown, neither the measurement nor the modeling error is known. However, we can try to describe them in statistical terms by assuming that they follow a certain distribution. Furthermore, the difference between the measured and the modeled quantities (i.e., the residuals) can be calculated. Note that the aim of inverse modeling is to provide an estimate of the true system behavior. If the true values is identified, the residuals are equal to the measurement errors. In other words, the statistical characteristic of the residuals should be similar to that of the measurement errors.

It is very important to appreciate the difference between systematic and random components in the residuals. The difference between systematic and random components and their relation to the functional and stochastic model, respectively, are illustrated in Table 2.1. The systematic component of the system response is hopefully identified by accurately modeling the physical behavior. This is referred to as the functional model which includes the conceptual model, the governing equations, the parameters entering these equations, and the numerical scheme used to solve the problem.

Provided that the true system behavior is identified, the residuals become equal to the random components of the observed system response, which are usually associated with the measurement errors. While the individual measurement errors are not known *a priori*, they can be described in

Figure 2.1: True, measured, and calculated system response, definition of measurement and modeling error.

Table 2.1: Systematic and Random Part of System Response

data	=	true response	+	measurement error
		identified component		unidentified component
		systematic		random
		conceptual		distributional assumption
		functional model		stochastic model
		TOUGH2		$\mathbf{C} = \sigma_0^2\mathbf{V}$
calibration point	=	fitted value	+	residual

statistical terms. The so-called stochastic model comprises our assumptions about the distribution of the measurement errors.

The assumption that the system response can be separated into a systematic part (modeled by the process simulator), and a random part (described by the stochastic model) requires that all systematic errors are removed from both the data and the model, i.e., the final residuals should only contain random components that are accurately described by the stochastic model.

By definition, outliers do not conform to the assumed distributions. Moreover, if systematic errors are present, the estimated parameters are likely to be biased. This bias is often significantly larger than parameter uncertainties that result from random measurement errors.

In the following paragraphs we discuss a few potential sources for systematic errors in the analysis of the gas-pressure-pulse-decay (GPPD) experiment introduced in Section 1.3. Systematic errors occur in both the data and the numerical simulation. In many cases it is difficult and also irrelevant to distinguish between a systematic modeling error and a systematic error in the data. Systematic errors are simply the result of a conceptual difference between the observation and the corresponding model output. It is more a question of convenience which side of the problem can be better controlled.

Data from laboratory experiments are almost always easier to invert than field data because the key processes involved are better understood, the flow geometry is well known, and initial and boundary conditions are better controlled. Nevertheless, a careful design of the testing apparatus is important. For example, the volumes of the upstream and downstream reservoir of the GPPD experiment have to be accurately determined; system compliance effects should be minimized by choosing appropriate equipment materials; leaking must be avoided by applying sufficient confining pressures; temperature should be kept constant.

Deviations from these conditions have to be corrected in the data, if possible, or accurately reproduced in the model. For example, if temperature varies during the course of the experiment, the pressure data can be adjusted according to the ideal gas law. Alternatively, one could use a nonisothermal model that directly accounts for the temperature dependency of density, viscosity, and Klinkenberg factor. Note that while the latter approach is more difficult to implement, it is also more accurate.

The TOUGH2 model used to analyze the GPPD data is based on an equation-of-state module that describes the thermophysical properties of air rather than nitrogen—the gas used in the experiment. Differences in density and viscosity affect the pressure transient and thus the estimates. Density-viscosity ratios between air and nitrogen differ by a factor of about 1.05. This leads to an underestimation of permeability by 5% if pressure data from an experiment with nitrogen are inversely analyzed using air properties. In this case, the discrepancy between the data and the model output can be compensated after the inversion. In most instances, however, when the model output is affected in a nonlinear fashion, such corrections are not possible. The estimation procedure must then be repeated with different assumptions regarding those aspects of the model that are considered uncertain. This may provide some insight into the sensitivity of the results with respect to individual errors. However, the impact of a combination of errors is difficult to assess. In some cases, potential errors can be parameterized and subjected to the estimation process. An example of this approach is discussed in [9], where uncertainties regarding initial conditions and potential leaking are addressed.

In summary, the stochastic model deals with the random errors, assuming that the systematic part of the system behavior is adequately reproduced by the simulation model.

2.2 Observation Covariance Matrix

In the previous section we saw that the unexplained part of the system response cannot be described individually, but by means of a stochastic model. We have to make an assumption about the distribution of the measurement errors. The distribution of the final residuals is supposed to be identical with the distribution of the measurement errors, assuming that the true system response is correctly identified by the model.

A reasonable assumption about the measurement errors is that they are uncorrelated, normally distributed random variables with zero mean (the residual analysis will have to show whether this assumption is justified). The distributional assumption can therefore be summarized in a covariance matrix $\mathbf{C}_{zz}$. $\mathbf{C}_{zz}$ is an $m \times m$ diagonal matrix. The i-th diagonal element of matrix $\mathbf{C}_{zz}$ is the variance representing the measurement error of observation z_i.

$$\mathbf{C}_{zz} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_i^2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_n^2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \sigma_j^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \sigma_m^2 \end{bmatrix} \tag{2.1}$$

The purpose and interpretation of the elements of $\mathbf{C}_{zz}$ are manifold:

- They scale data of different quality, i.e., an accurate measurement obtains a higher weight than a more uncertain measurement.
- They scale observations of different types. For example, flow rates and pressures have different units and usually differ by many orders of magnitudes. They need to be scaled appropriately to be comparable in a formalized parameter estimation procedure.
- They weight the fitting error.
- $\mathbf{C}_{zz}$ is the stochastic model for maximum-likelihood estimation for normally distributed residuals.

One should realize that only the ratio σ_i^2/σ_j^2 is important, i.e., the estimated parameter set is not affected by a linear scaling of the covariance matrix. We can therefore introduce a factor σ_0^2 and write:

$$\mathbf{C}_{zz} = \sigma_0^2 \cdot \mathbf{V}_{zz} \tag{2.2}$$

where $\mathbf{V}_{zz}$ is a positive definite matrix. σ_0^2 is termed the *a priori* error variance. It is the variance of a dimensionless error of size one. Since it can assume any positive value, we select it to be equal to 1.0 and directly work with the actual covariance matrix rather than $\mathbf{V}_{zz}$. After the inversion, the *a posteriori* or estimated error variance s_0^2 is calculated. If the assumption about the measurement errors was correct, and if the true system response is identified, the ratio s_0^2/σ_0^2 will not significantly deviate from 1.0 (Section 4.1.2 contains more details about the Fisher model test).

2.3 Objective Function

2.3.1 The norm as a measure of misfit

The objective function is a measure of the misfit between the data and the model calculation. It is also termed performance measure, penalty function, energy function, norm, misfit criterion, etc.

There are many ways to measure the difference between the observed and calculated system response. In the standard trial-and-error calibration procedure, the simulation results and data are plotted, and a rather subjective judgment is made as to how well the calculation matches the data. A more objective way is to calculate a norm of the residual vector. A norm of a vector is defined as follows:

$$\|\mathbf{r}\| = \left(\sum_{i=1}^{m} |r_i|^p \right)^{1/p} \tag{2.3}$$

The most common norms are the L_1-norm, the L_2- or Euclidean norm, and L_∞, leading to the L_1-estimator, Least Squares, and Minmax, respectively.

The maximum likelihood approach discussed in Section 2.3.2 takes the distributional assumption about the measurement errors as a basis for choosing the objective function. It can be shown that normally distributed errors lead to the well-known generalized least-squares objective function. The central limit theorem makes least-squares a reasonable choice. However, the distribution of the residuals often deviates from being Gaussian. For example, the presence of outliers in the data or systematic modeling errors lead to nonsymmetric distributions with stronger tails than predicted by the normal distribution. For these cases, alternative objective functions may be more appropriate to avoid biased estimates. These so-called *robust estimators* are discussed in [10].

The minimum of the objective function indicates the best attainable fit to the data. The parameters at the minimum are therefore considered *best estimates*.

2.3.2 Maximum likelihood

Let $\mathbf{p}$ be the parameter vector of length n, an $\mathbf{z}$ the observation vector of length m. The joint probability density function (pdf), $\Phi(\mathbf{z};\mathbf{p})$, is defined as the probability of observing the data $\mathbf{z}^*$ if $\mathbf{p}$ were true: $\Phi(\mathbf{z};\mathbf{p}) =$

$\Pr(\mathbf{z} = \mathbf{z}^*|\mathbf{p})$. Note that $\mathbf{p}$ is unknown, nevertheless deterministic. If the observations are independent random variables, the joint pdf is given by the product of the probabilities of the individual observations:

$$\Phi(\mathbf{z}; \mathbf{p}) = \prod_{i=1}^{n} \Phi(z_i; \mathbf{p}) \tag{2.4}$$

From a different perspective, this equation may be seen as describing the likelihood of $\mathbf{p}$ if the values z_i^* are fixed. This is termed the *likelihood function*:

$$\Phi(\mathbf{z}; \mathbf{p}) \Leftrightarrow L(\mathbf{p}; \mathbf{z}^*) \tag{2.5}$$

For each parameter set $\mathbf{p}$, the likelihood function $L(\mathbf{p}; \mathbf{z}^*)$ gives the probability of observing $\mathbf{z}^*$. Thus we can think of $L(\mathbf{p}; \mathbf{z}^*)$ as a measure of how 'likely' $\mathbf{p}$ is to have produced the observed data $\mathbf{z}^*$. The method of maximum likelihood consists of finding the specific value of the parameter that is 'most likely' to have produced the data.

The maximum likelihood approach is discussed for the normal distribution in the following section.

2.3.3 Least squares

Let's assume that all measurement errors $(\mathbf{z}^* - \hat{\mathbf{z}})$ are normally distributed with mean $E[(\mathbf{z}^* - \hat{\mathbf{z}})] = 0$ and covariance matrix $E[(\mathbf{z}^* - \hat{\mathbf{z}})(\mathbf{z}^* - \hat{\mathbf{z}})^T] = \mathbf{C}_{zz}$ (if the errors follow a log-normal distribution, they have to be transformed by taking the logarithm of $(z_i^* - \hat{z}_i)$ to yield Gaussian distributions). Note that the true, albeit unknown value of z is denoted by $\hat{z}$. The likelihood function can be written as follows:

$$L(\mathbf{p}; \mathbf{z}^*) = (2\pi)^{-m/2} \left|\mathbf{C}_{zz}\right|^{-1/2} \cdot \exp\left[-\frac{1}{2}(\mathbf{z}^* - \hat{\mathbf{z}})^{\mathbf{T}} \, \mathbf{C}_{zz}^{-1} \, (\mathbf{z}^* - \hat{\mathbf{z}})\right] \tag{2.6}$$

It is obvious that if the stochastic model—the covariance matrix $\mathbf{C}_{zz}$—is known, maximizing (2.6) is equivalent to minimizing the objective function

$$S = (\mathbf{z}^* - \mathbf{z}(\mathbf{p}))^{\mathbf{T}} \, \mathbf{C}_{zz}^{-1} \, (\mathbf{z}^* - \mathbf{z}(\mathbf{p})) \tag{2.7}$$

Note that we have replaced the true value $\hat{z}$ with the calculated value $\mathbf{z}(\mathbf{p})$. Since $(\mathbf{z}^* - \mathbf{z}(\mathbf{p})) = \mathbf{r}$, and $\mathbf{C}_{zz}$ is a diagonal matrix, S is the sum of the squared residuals, weighted by the inverse of the prior variances σ_i^2:

$$S = \sum_{i=1}^{m} \frac{r_i^2}{\sigma_i^2} \tag{2.8}$$

In summary: If the errors are normally distributed, the method of least squares leads to maximum likelihood parameter estimates. The best estimate is the parameter vector $\mathbf{p}$ that minimizes the sum of the squared, weighted residuals.

2.4 Summary

In inverse modeling, parameters are derived from data that consist of two components, (1) a *systematic* component reflecting the system behavior, and (2) a *random* component stemming from unexplained measurement errors. The functional model deals with the systematic component, whereas the stochastic model describes the random errors. If the distribution of the final residuals after calibration is not consistent with the assumed distribution of the measurement errors, this indicates that there is either an error in the functional model, a systematic error in the data, or an error in the stochastic model.

In *maximum likelihood* estimation one tries to find the parameters that maximize the likelihood of the model to produce the observed data. Minimizing the sum of the weighted squared residuals leads to maximum likelihood estimates provided that the residuals follow a normal distribution.

The *objective function* represents the measure of misfit between the data and the model calculation. Parameters are estimated by iteratively minimizing the objective function.

Chapter 3

Minimization Algorithm

3.1 Purpose and Classification

The purpose of the minimization algorithm is to find the minimum of the objective function by iteratively updating the parameters of the model. The search for the minimum occures in the n-dimensional parameter space. Recall the the objective function is a global measure of the misfit between the data and the corresponding model output. Since the model output $z_i(\mathbf{p})$ depends on the parameters, the fit can be improved by changing the elements of the parameter vector $\mathbf{p}$. There are a number of strategies to find parameter combinations that yield smaller values of the objective function, eventually identifying a local or hopefully global minimum. The available methods can be classified as follows:

Direct Search Methods In these methods the objective function is evaluated for different parameter combinations, mapping out the objective function in the n-dimensional parameter space, looking for the minimum. While no derivatives of the objective functions with respect to the parameters must be calculated, these methods usually require many function evaluations (i.e., solutions of the forward problem) and are therefore inefficient. Examples of Direct Search Methods include:

- Trial-and-error model calibration
- Grid search
- Simplex algorithm
- Simulated annealing
- Genetic algorithms
- ...

Gradient-Based Methods These methods require calculating the gradient of the objective function with respect to the parameter vector. Updating the parameter vector in small steps along the direction given by the gradient is a robust albeit inefficient procedure. Various modifications of that basic scheme have been proposed. Examples of gradient-based minimization algorithms include:

- Steepest descent
- Conjugate gradient method
- ...

Second-Order Methods These methods are based on the Hessian matrix or various approximations thereof. The computational cost for calculating the second derivatives is compensated by a rather efficient stepping in the parameter space. Examples of second-order methods include:

- Newton method
- Gauss-Newton method
- Levenberg-Marquardt
- ...

Other classifications have been proposed in the literature [12]. Each of the methods mentioned above has its advantages and disadvantages. In the remainder of this chapter we focus on the Levenberg-Marquardt modification of the Gauss-Newton algorithm for minimizing objective functions from highly nonlinear models.

3.2 Gauss-Newton

We discuss the Gauss-Newton method for the minimization of the least-squares objective function

$$S = (\mathbf{z}^* - \mathbf{z}(\mathbf{p}))^T \, \mathbf{C}_{zz}^{-1} \, (\mathbf{z}^* - \mathbf{z}(\mathbf{p})) = \mathbf{r}^T \mathbf{C}_{zz}^{-1} \mathbf{r} \tag{3.1}$$

Let's assume for the moment that the model is linear in the parameters, making the objective function (3.1) quadratic as illustrated for one and two parameters in Figure 3.1.

In multiphase flow modeling, the discretized flow and transport equations are highly nonlinear functions of the parameters to be estimated. If the model is nonlinear, the objective function is no longer quadratic. However, most minimization algorithms for least-squares problems rely on a local quadratic approximation to the objective function as shown in Figure 3.2.

The Gauss-Newton method can be derived directly from the objective function. If the model is linear, the Gauss-Newton method leads to the optimum parameter set in a single step. If the model is nonlinear, the objective function is approximated at the current point in the parameter space, and

Figure 3.1: Objective function for a linear model as a function of (a) one and (b) two parameters.

a Gauss-Newton step is performed as outlined below. The procedure is repeated from the newly obtained parameter set, until a convergence criterion is met.

Local linearization means that we take the partial derivatives of the model output at the calibration points—the elements of vector $\mathbf{z}$—with respect to the parameters of interest—the elements of vector $\mathbf{p}$. This yields the so-called *Jacobian* matrix $\mathbf{J}$ of dimensions $m \times n$, the elements of which are defined as

$$J_{ij} = \frac{\partial z_i}{\partial p_j} \qquad i = 1, \ldots, m \qquad j = 1, \ldots, n \tag{3.2}$$

There are several methods to calculate the elements J_{ij}. The simplest, albeit computationally costly way is the perturbation method using either forward or centered finite differences:

$$\text{forward:} \qquad J_{ij} \approx \frac{z_i(\mathbf{p}; p_j + \delta p_j) - z_i(\mathbf{p})}{\delta p_j} \tag{3.3}$$

$$\text{centered:} \qquad J_{ij} \approx \frac{z_i(\mathbf{p}; p_j + \delta p_j) - z_i(\mathbf{p}; p_j - \delta p_j)}{2\delta p_j} \tag{3.4}$$

where each parameter is perturbed by a small percentage (e.g., $\alpha \approx 0.01$) of its value:

$$\delta p_j = \alpha p_j \tag{3.5}$$

Figure 3.2: Objective function for a nonlinear model as a function of (a) one and (b) two parameters. Also shown are the quadratic approximations at $\mathbf{p}^*$ and Gauss-Newton steps.

Note that calculating a forward finite difference approximation of the Jacobian matrix $\mathbf{J}$ requires $n+1$ solutions of the forward problem, i.e., $n+1$ transient simulations from time zero to the time of the last calibration point. Centered finite differences (3.4) are more accurate, but require $2n+1$ runs. In most cases, high accuracy is not essential far away from the minimum, i.e., at the beginning of the optimization process, but may become crucial as the minimum is approached. Furthermore, the linear error analysis discussed in Chapter 4 depends on the Jacobian evaluated at the minimum.

The Gauss-Newton solution in the linear case can easily be derived from the objective function (3.1) by first noting that $\mathbf{z} = \mathbf{J}\mathbf{p}$,

$$S = (\mathbf{z}^* - \mathbf{J}\mathbf{p})^T \mathbf{C}_{zz}^{-1} (\mathbf{z}^* - \mathbf{J}\mathbf{p}) \tag{3.6}$$

and then setting the derivative to zero in order to obtain the minimum:

$$\begin{aligned}
\frac{\partial S}{\partial \mathbf{p}} &= (-\mathbf{J})^T \mathbf{C}_{zz}^{-1} (\mathbf{z}^* - \mathbf{J}\mathbf{p}) + (\mathbf{z}^* - \mathbf{J}\mathbf{p})^T \mathbf{C}_{zz}^{-1} (-\mathbf{J}) \\
&= -\mathbf{J}^T \mathbf{C}_{zz}^{-1} \mathbf{z}^* + \mathbf{J}^T \mathbf{C}_{zz}^{-1} \mathbf{J}\mathbf{p} - \mathbf{z}^{*T} \mathbf{C}_{zz}^{-1} \mathbf{J} + (\mathbf{J}\mathbf{p})^T \mathbf{C}_{zz}^{-1} \mathbf{J}
\end{aligned}$$

Table 3.1: The Gauss-Newton Minimization Algorithm

Step 1:	Initialization: • Set iteration index $k = 0$. • Define initial parameter set $\mathbf{p}^{(k=0)}$.
Step 2:	Run simulation model with parameter vector $\mathbf{p}^{(k)}$.
Step 3:	Evaluate $\mathbf{r}(\mathbf{p}^{(k)})$, $S(\mathbf{p}^{(k)})$, and $\mathbf{J}(\mathbf{p}^{(k)})$.
Step 4:	Calculate parameter update: $\Delta\mathbf{p} = \left(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{r}$
Step 5:	Update parameter vector: $\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} + \Delta\mathbf{p}$
Step 6:	Evaluate $S(\mathbf{p}^{(k+1)})$.
Step 7:	Evaluate convergence criteria. If converged, go to Step 8, else set $k = k + 1$ and go to Step 2.
Step 8:	Minimum identified. Proceed with error analysis.

$$= \quad 2\left(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J}\mathbf{p} - \mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{z}^*\right) = 0 \tag{3.7}$$

From this we obtain the solution vector $\mathbf{p}$ that minimizes S:

$$\mathbf{p} = \left(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{z}^* \tag{3.8}$$

Matrix $(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J})$ is symmetric and of dimension $n \times n$. It is sometimes termed *Fisher Information Matrix.* The Fisher information matrix is an approximation of the Hessian matrix of S.

The best estimate parameter set is odirectly btained only if the model is linear. For nonlinear models, the solution has to be sought iteratively. In the iterative scheme, the data vector $\mathbf{z}^*$ is replaced by the residual vector $\mathbf{r}(\mathbf{p}^{(k)})$, where the superscript $^{(k)}$ labels the k-th iteration or minimization step. Then, the vector $\Delta\mathbf{p}$ holding the parameter updates can be calculated. The Gauss-Newton method is summarized in Table 3.1.

The Gauss-Newton step is very efficient if the model is linear or nearly linear, i.e., if $(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J})$ is a good approximation of the Hessian. If the model is highly nonlinear, however, the parameter update (Step 4 in Table 3.1) can be too large, leading to an inefficient or even unsuccessful step where the value of the objective function S is increased rather than decreased.

3.3 Levenberg-Marquardt

For nonlinear models and if the parameter vector $\mathbf{p}$ is far away from the optimum parameter set, the Hessian is not necessarily a positive-definite matrix, and its approximation by matrix $(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J})$ (see Gauss-Newton method, Section 3.2) may not lead to a successful or efficient step. Recall that the Hessian is an $n \times n$ matrix with the second partial derivatives of the objective function $S = \mathbf{r}^T\mathbf{C}_{zz}^{-1}\mathbf{r}$. Its elements can be written as follows:

$$H_{jk} = 2\sum_{i=1}^{m}\left[\frac{1}{\sigma_i^2}\left(\frac{\partial r_i}{\partial p_j}\frac{\partial r_i}{\partial p_k} + r_i\frac{\partial^2 r_i}{\partial p_j \partial p_k}\right)\right] \tag{3.9}$$

In matrix form, the Hessian reads

$$\mathbf{H} = 2\left(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J} + \sum_{i=1}^{m} r_i\mathbf{G}_i\right) \tag{3.10}$$

where $\mathbf{G}_i = \nabla^2 r_i/\sigma_i$ is the Hessian of the weighted residuals. Denoting the sum in Equation (3.10) with $\mathbf{S}$, the Hessian becomes

$$\mathbf{H} = 2\left(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J} + \mathbf{S}\right) \tag{3.11}$$

Note that $\mathbf{S}$ is zero if the model is linear, confirming the solution previously discussed in Section 3.2. However, $\mathbf{S}$ cannot be neglected in the nonlinear case, especially if the residuals are large, i.e., far away from the minimum. Also note that the positive and negative residuals do not cancel one another, i.e., the Hessian is not necessarily a positive-definite matrix.

The various iterative solutions to the nonlinear least-squares problem are based on different approximations to the Hessian. While the first term in Equation (3.11) is relatively easy to calculate, the evaluation of $\mathbf{S}$ is computationally costly. However, it may not be necessary to evaluate $\mathbf{S}$ if it is small compared with $(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J})$. As discussed above, the Gauss-Newton method simply ignores $\mathbf{S}$, assuming that the model is only slightly nonlinear and that the residuals are relatively small as is the case close to the minimum.

In the *Levenberg-Marquardt* method, the approximation to the Hessian is assured to be positive-definite by replacing $\mathbf{S}$ with an $n \times n$ diagonal matrix $\lambda\mathbf{D}$. The scalar λ is the so-called *Levenberg parameter*. It is updated following a strategy proposed by *Marquardt* [16]. The Levenberg-Marquardt minimization algorithm is described in Table 3.2.

Far away from the minimum, i.e., during the first few iterations, a relatively large value of λ is chosen, leading to a small step along the gradient of S. Stepping along the steepest descent direction is robust, but inefficient. The Levenberg parameter is decreased after each successful step. With decreasing λ, as the minimum is approached, the parameter update

Figure 3.3: Steps proposed by the Levenberg-Marquardt method as a function of the Levenberg parameter λ.

$\Delta\mathbf{p}$ approximates that proposed by the Gauss-Newton algorithm with its quadratic convergence rate. Figure 3.3 shows various end points of steps taken by the Levenberg-Marquardt algorithm as a function of λ.

3.4 Stopping Criteria

As we have seen in the previous sections, the minimum is approached by proposing new parameter sets that lead to reduced values of the objective function. Stopping criteria have to be specified to decide whether the minimum has been identified. Theroetically, the minimum of the objective function is detected if all the elements of the gradient vector $\partial S/\partial\mathbf{p}$ are zero. In practice, however, one of the following convergence criteria are used to stop optimization:

- The objective function is smaller than a specified tolerance;
- The normalized step size is smaller than a minimum relative step size;
- The norm of the gradient vector is smaller than a specified tolerance;
- The Levenberg parameter exceeds a specified value;
- The number of unsuccessful uphill step exceeds a specified tolerance;
- All parameters are at their lower or upper bounds;
- The number of iteration exceeds a specified value;

- ...

Note that the convergence criteria may vary if a minimization algorithm other than Levenberg-Marquardt is employed. For example, no gradient is available when using one of the Direct Search Methods, making it impossible to use the optimality measure as a stopping criterion.

The objective function is usually substantially reduced during the first few optimization steps. Limiting the number of iterations based on experience is thus a reasonable convergence criterion.

Table 3.2: The Levenberg-Marquardt Minimization Algorithm

Step 1:	Initialization: • Set iteration index $k = 0$. • Define initial value of Levenberg parameter λ. • Define Marquardt parameter $\nu > 1$. • Define initial parameter set $\mathbf{p}^{(k=0)}$.
Step 2:	Run simulation model with parameter vector $\mathbf{p}^{(k)}$.
Step 3:	Evaluate $\mathbf{r}(\mathbf{p}^{(k)})$, $S(\mathbf{p}^{(k)})$, and $\mathbf{J}(\mathbf{p}^{(k)})$.
Step 4:	Propose parameter update: $\Delta\mathbf{p} = \left(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J} + \lambda\mathbf{D}\right)^{-1}\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{r}$ where $\mathbf{D}$ is an $n \times n$ diagonal matrix with $D_{ii} = (\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J})_{ii}$
Step 5:	Update parameter vector: $\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} + \Delta\mathbf{p}$
Step 6:	Evaluate $S(\mathbf{p}^{(k+1)})$.
Step 7:	If $S(\mathbf{p}^{(k+1)}) \geq S(\mathbf{p}^{(k)})$ multiply λ by ν and go to Step 4. If $S(\mathbf{p}^{(k+1)}) < S(\mathbf{p}^{(k)})$ divide λ by ν and go to Step 8.
Step 8:	Evaluate convergence criteria. If converged, go to Step 9, else set $k = k + 1$ and go to Step 2.
Step 9:	Minimum identified. Proceed with error analysis.

Chapter 4

Error Analysis

4.1 *A Posteriori* Error Analysis

4.1.1 Introduction

One of the key advantages of a formalized approach to parameter estimation is the possibility to perform an extensive *a posteriori* error analysis. First, the residual analysis provides some measure of the overall goodness-of-fit, identifies systematic errors, trends in the model, or outliers in the data. Next we can determine the uncertainty of the estimated parameters. Note that a good match does not necessarily mean that the estimates are reasonable. They may be highly uncertain due to high parameter correlation (an indication of overparameterization). The covariance matrix of the estimated parameters can be further analyzed to obtain correlation coefficients, parameter combinations that lead to similar matches, etc. Finally, we can calculate the uncertainty of the model predictions using either linear error propagation analysis or Monte Carlo simulations.

4.1.2 Estimated error variance s_0^2 and Fisher Model Test

The estimated error variance represents the variance of the mean weighted residual and is thus a measure of goodness-of-fit:

$$s_0^2 = \frac{\mathbf{r}^T \mathbf{C}_{zz}^{-1} \mathbf{r}}{m - n} \tag{4.1}$$

Note that if the residuals are consistent with the distributional assumption about the measurement errors (covariance matrix $\mathbf{C}_{zz}$), then the estimated error variance assumes a value close to one. Since s_0^2 is an estimate of σ_0^2 (see Equation (2.2)), the denominator is $(m - n)$ rather than just m. The difference between the number of calibration points and the number of parameters $(m - n)$ is called the *degree of freedom.* It can be shown that the ratio s_0^2/σ_0^2 follows an F-distribution with the two degrees of freedom

Table 4.1: Fisher Model Test

$F_{m-n.\infty,1-\alpha} < s_0^2/\sigma_0^2$	⇒	Error in functional or stochastic model.
$1 \leq s_0^2/\sigma_0^2 \leq F_{m-n.\infty,1-\alpha}$	⇒	Model test passed; use s_0^2 for subsequent error analysis.
$s_0^2/\sigma_0^2 < 1$	⇒	Error in stochastic model; use σ_0^2 for subsequent error analysis.

$f_1 = (m - n)$, and $f_2 = \infty$. We can therefore statistically test whether the match deviates significantly from the modeler's expectations which were expressed through matrix $\mathbf{C}_{zz}$. The *Fisher model test* is shown in Table 4.1.

Note that the Fisher model test is only useful if the stochastic model is well defined. Otherwise, the ratio s_0^2/σ_0^2 is just a relative measure of goodness-of-fit.

4.1.3 Covariance matrix of estimated parameters

Next we calculate the expected value and the covariance matrix of the estimated parameters, $\mathbf{C}_{pp}$. Based on the linearity assumption we obtain for the estimated parameter vector (see also Equation (3.8)):

$$\hat{\mathbf{p}} = \left(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{z}^* \tag{4.2}$$

Evaluating the expected value of $\hat{\mathbf{p}}$ yields:

$$\mathrm{E}[\hat{\mathbf{p}}] = \left(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^T\mathbf{C}_{zz}^{-1}\underbrace{\mathrm{E}[\mathbf{z}^*]}_{\mathbf{J}\mathbf{p}} = \mathbf{p} \tag{4.3}$$

Equation(4.3) demonstrates that the expected value of the estimated parameter is equal to the parameter itself, i.e., (4.2) is an *unbiased* estimator.

The covariance matrix is defined as

$$\mathbf{C}_{pp} = \mathrm{E}[(\hat{\mathbf{p}} - \mathrm{E}[\hat{\mathbf{p}}])(\hat{\mathbf{p}} - \mathrm{E}[\hat{\mathbf{p}}])^T] \tag{4.4}$$

Inserting Equation (4.2) for $\hat{\mathbf{p}}$ and Equation (4.3) for $\mathrm{E}[\hat{\mathbf{p}}]$ yields after some rearrangement:

$$\mathbf{C}_{pp} = s_0^2 \cdot \left(\mathbf{J}^T\mathbf{C}_{zz}^{-1}\mathbf{J}\right)^{-1} \tag{4.5}$$

The Jacobian $\mathbf{J}$ is evaluated at the optimum parameter set. The interpretation of the covariance matrix (4.5) provides the key criteria to evaluate inverse modeling results. First we note that $\mathbf{C}_{pp}$ is directly proportional to the overall goodness-of-fit expressed by s_0^2. The diagonal elements of $\mathbf{C}_{pp}$ contain the variances σ_{ii}^2 of the estimated parameters.

Next we shortly discuss the impact of correlations on the estimation error. Correlations among parameters can be described as a conjoint impact of parameter changes on the system behavior. The correlation coefficient is given by:

$$r_{ij} = \frac{c_{ij}}{\sqrt{\sigma_{ii}^2 \cdot \sigma_{jj}^2}} \tag{4.6}$$

where c_{ij} are the covariances, i.e., the off-diagonal elements of $\mathbf{C}_{pp}$. The correlation coefficient assumes values between -1 and 1; a value of zero indicates no statistical correlation between parameter i and j, a value close to -1 or 1 indicates a strong correlation, i.e., the two parameters cannot be determined independently. For example, if two parameters are negatively correlated, a similar system response is obtained by concurrently increasing one and decreasing the other parameter. Even though certain pairs of parameters may exhibit preferential correlation structures, correlations are not invariable entities of parameter combinations. They obviously depend on the data available, and also on the number of simultaneously estimated parameters, since indirect correlations may overwhelm the direct correlations. If correlations exist, the uncertainty of one parameter does affect the uncertainty of the other parameter. The diagonal elements of matrix $\mathbf{C}_{pp}$, which are the variances from the joint probability density function, account for this fact. They have to be distinguished from the conditional standard deviation σ_{ii}^* which measures the uncertainty of a parameter assuming that all the other parameters are either exactly known or uncorrelated. The conditional standard deviation is obviously always smaller than that from the joint probability density function. The situation is illustrated in Figure 4.1 for the case of two parameters. The ratio

$$\chi_i = \frac{\sigma_{ii}^*}{\sigma_{ii}} \tag{4.7}$$

is a measure of how independently parameter i can be estimated. Small values of χ_i usually indicate that the uncertainty σ_{ii} of a parameter could be reduced by lowering its correlation to other parameters. We mention in passing that the length and orientation of the semiaxis of the elliptical confidence region shown in Figure 4.1 can be obtained from an eigenanalysis of matrix $\mathbf{C}_{pp}$.

An example illustrates the importance of the error analysis in inverse modeling. Table 4.2 shows the covariance matrix from the GPPD experiment discussed in Section 1.3. The diagonal contains the variances, the lower triangle is the covariance matrix, and the upper triangle holds the corresponding correlation coefficients calculated using Equation(4.6).

From the perfect match (see Figure 1.2) and generally high sensitivity coefficients, one might expect that an accurate estimation of the three parameters is possible. However, an inspection of the covariance matrix in

Figure 4.1: Two-dimensional confidence region from linear error analysis. Joint and conditional standard deviations are indicated.

Table 4.2: Estimation Covariance and Correlation Matrices from GPPD Experiment

	$\log(k)$	$\log(b)$	Porosity	$\sigma^*_{ii}/\sigma_{ii}$
$\log(k)$	1.67	< -0.99	-0.87	< 0.01
$\log(b)$	-1.90	2.16	0.87	< 0.01
Porosity	-5.70×10^{-4}	6.59×10^{-4}	2.64×10^{-7}	0.48

Table 4.2 reveals a large estimation uncertainty. The standard deviation of both permeability and Klinkenberg factor is greater than one order of magnitude. This is obviously a result of the high correlation between the two parameters. The correlation coefficient between $\log(k)$ and $\log(b)$ is very close to -1, indicating that an increase in one parameter can be almost completely compensated by a decrease in the other parameter. The physical explanation is evident from Equation (1.1), where k and b become linearly dependent for a constant average pressure within the sample. The ratio of the conditional and joint standard deviation shown in the last column of Table 4.2 confirms the high dependency between the two parameters $\log(k)$ and $\log(b)$.

The pressure dependency of gas slip flow suggests that the statistical correlation between the two parameters of interest can be reduced by performing experiments at different pressure levels. A simultaneous inversion of data from several experiments should yield a unique and stable solution with low correlation coefficients and low estimation uncertainties. Data from three GPPD experiments were simultaneously matched as shown in Figure 4.2.

Table 4.3 shows that the correlation between $\log(k)$ and $\log(b)$ is weakened from -0.99 in the previous case to -0.52. As expected, this leads to more independent estimates as implied by the values in the last column of

Table 4.3: Estimation Covariance and Correlation Matrices from Modified GPPD Experiment

	$\log(k)$	$\log(b)$	Porosity	$\sigma^*_{ii}/\sigma_{ii}$
$\log(k)$	1.04×10^{-4}	-0.52	-0.12	0.85
$\log(b)$	-1.07×10^{-4}	4.10×10^{-4}	-0.02	0.85
Porosity	-1.30×10^{-6}	-3.62×10^{-7}	1.06×10^{-6}	0.99

Table 4.3, and a significant reduction in the estimation error.

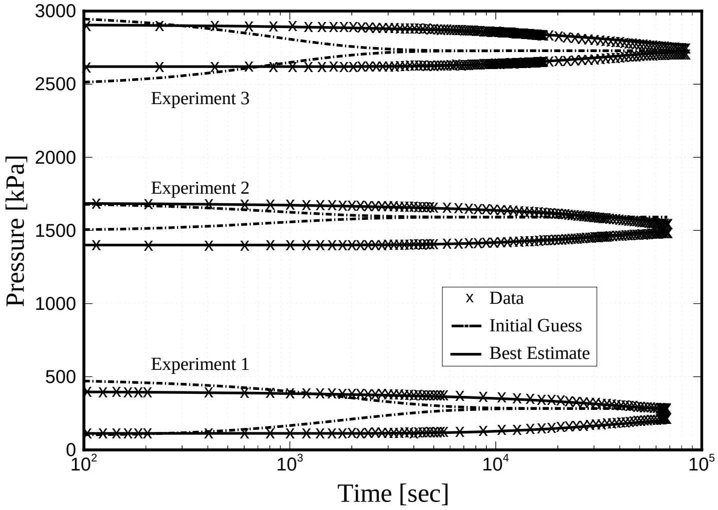

Figure 4.2: Comparison between measured and calculated pressure transient curves with the initial and final parameter set.

This example demonstrates that a good match and high parameter sensitivities are not sufficient to guarantee a meaningful solution of the inverse problem. The *a posteriori* error analysis first revealed high estimation uncertainties, and suggested modifying the experiment such that the correlation between $\log(k)$ and $\log(b)$ be reduced. As a result, a successful inversion was performed leading to accurate estimates of the parameters of interest.

4.1.4 Residual analysis and model identification

As mentioned earlier, maximum likelihood estimation leads to optimum parameters for a given model structure. However, this does not imply that the representation of the real system is satisfying. If the conceptual model fails to reproduce the salient features of the system, the calibrated model may not be able to match the observed data as expected (recall that our expectation regarding the fit is reflected in the *a priori* covariance matrix $\mathbf{C}_{zz}$). The Fisher Model Test outlined in Section 4.1.2 is a first indication of whether the model fits the data well enough so that the underlying conceptual model can be accepted. Furthermore, a detailed residual analysis may reveal trends in the residuals, indicating that there is a systematic error in the model or the data.

Figure 4.3 shows the residuals from the simultaneous inversion of three GPPD experiments. The apparently good match depicted in Figure 4.2 in fact exhibits a systematic error at late times for Experiments No. 2 and 3, which were performed at higher pressure levels. The fact that the model systematically overpredicts the measured data suggests that there is a gas leak in the experimental apparatus. The gas leak can be incorporated into the numerical model by specifying a sink term. Its value can be estimated along with the other parameters. Figure 4.4 shows that the systematic errors vanish by estimating two sink terms representing the gas leaks. This leads to randomly distributed residuals consistent with the assumptions described in the stochastic model.

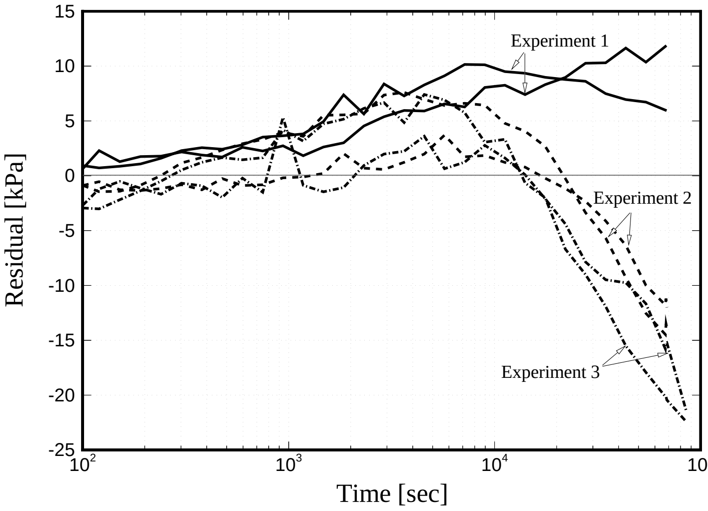

Figure 4.3: Residuals as a function of time, showing systematic overprediction of pressures at late times.

The desire to obtain a good match between the observed and predicted system response may tempt the modeler to increase the number of unknown parameters, as was done in the previous example, where two additional sink term parameters were introduced. Unfortunately, increasing the number of parameters results in a decrease in parameter reliability because the parameters become more strongly correlated. Furthermore, the degree of freedom is reduced. As a result, the model may become overparameterized.

The error analysis may indicate that too many parameters are included in the inversion. In addition, *model identification criteria* can be evaluated and used for selecting the most appropriate model. For more details about

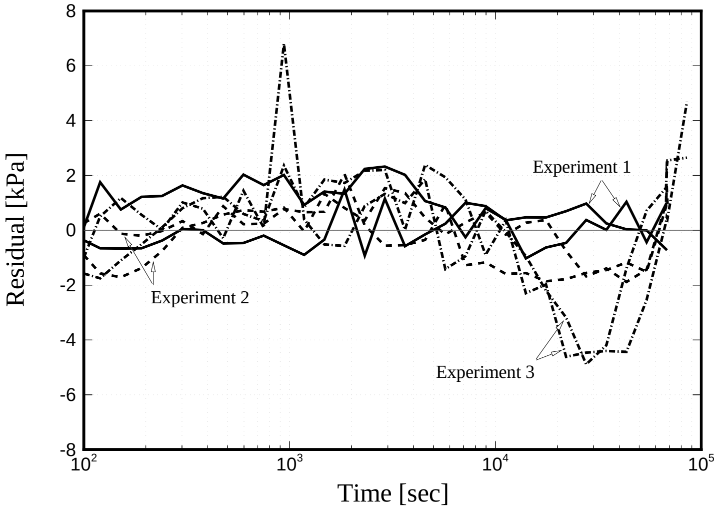

Figure 4.4: Residuals as a function of time after removal of systematic error by estimating sink terms (note the different scale).

this important topic, the reader is referred to [2][3][4][9].

4.2 Uncertainty Propagation Analysis

4.2.1 Introduction

Model predictions are inherently uncertain and may significantly deviate from the true system behavior. There are many reasons for the inconsistency between model predictions and the actual or observed system behavior. The main sources for modeling errors include:

- Inconsistencies and errors in the conceptual model;
- Uncertainty in the input parameters;
- Discretization errors.

As mentioned in Section 1.1, the conceptual model is by far the most important element in numerical modeling. Considerable effort should be spent on carefully developing the conceptual model, because errors in the model structure are difficult to identify and to correct, and they usually have the largest impact on the model predictions.

The second source of prediction errors is insufficient knowledge about the model parameters. Errors or uncertainties in the input parameters lead to errors or uncertainties in the model predictions. The purpose of inverse modeling is to estimate the best parameters for a given model structure, and to reduce parameter uncertainty. Nevertheless, there is a need to quantifiy the uncertainty in the model predictions as a result of parameter uncertainty, which is the topic of this lecture.

Finally, the numerical solution of the governing equation has only finite precision and may suffer from discretization errors such as numerical dispersion. While care must be taken when choosing the numerical scheme, errors from the numerical model are usually smaller than errors made by using wrong parameter values, which in turn are small compared with the errors from using an inappropriate conceptual model. One should also keep in mind that modeling involves an abstraction process, i.e., no exact solution is sought, but an approximation that is reasonable and capable of reproducing the salient features of the system to be studied.

4.2.2 Linear analysis

Linear or First-Order-Second-Moment (FOSM) uncertainty propagation analysis quantifies the uncertainty in model predictions as a result of parameter uncertainty. As the name indicates, FOSM is the analysis of the mean and covariance of a random function based on its first-order Taylor series expansion. The covariance of parameter estimates is translated into the covariance of the simulated system response. FOSM analysis presumes that the mean and covariance are sufficient to characterize the distribution of the dependent variables, i.e., the model results are assumed to be normally distributed. This assumption is valid whenever parameter uncertainties are sufficiently small, or when the model is linear and the distribution of the input parameters is normal. The normality and linearity assumptions must be checked before applying FOSM.

We develop expressions for the covariance matrix of the model prediction using first-order Taylor series expansion. Let $\hat{\mathbf{p}}$ be a vector of length n, holding the parameters considered uncertain. Its covariance matrix of dimension $n \times n$ is denoted by $\mathbf{C}_{pp}$. Furthermore, $\mathbf{z}$ is a vector of length m containing the simulation results at certain points in space and time. These model predictions are a function of the parameter vector $\mathbf{p}$. Finally, let $\mathbf{J}$ be the $m \times n$ Jacobian matrix holding sensitivity coefficients, i.e., $J_{ij} = \partial z_i / \partial p_j$. The model prediction $\mathbf{z}(\mathbf{p})$ can be approximated using first-order Taylor series expansion as follows:

$$\mathbf{z}(\mathbf{p}) \approx \mathbf{z}(\hat{\mathbf{p}}) + \mathbf{J}(\mathbf{p} - \hat{\mathbf{p}}) \tag{4.8}$$

The mean is given by:

$$\begin{aligned} \mathrm{E}[\mathbf{z}(\mathbf{p})] &\approx \mathrm{E}[\mathbf{z}(\hat{\mathbf{p}})] + \mathrm{E}[\mathbf{J}(\mathbf{p} - \hat{\mathbf{p}})] \\ &\approx \underbrace{\mathrm{E}[\mathbf{z}(\hat{\mathbf{p}})]}_{\mathbf{z}(\hat{\mathbf{p}})} + \mathrm{E}[\mathbf{J}] \cdot \underbrace{\mathrm{E}[(\mathbf{p} - \hat{\mathbf{p}})]}_{\mathbf{0}} \end{aligned}$$

$$\approx \mathbf{z}(\hat{\mathbf{p}}) \qquad (4.9)$$

The first-order approximation of the expected values of the dependent variables is the vector of the model prediction obtained using the mean parameters.

The covariance matrix of the simulated system response is derived using the following definition:

$$\begin{aligned}
\mathrm{Cov}[\mathbf{z}] &\approx \mathrm{E}[(\mathbf{z}-\hat{\mathbf{z}})(\mathbf{z}-\hat{\mathbf{z}})^T] && (4.10)\\
\mathrm{Cov}[\mathbf{z}(\mathbf{p})] &\approx \mathrm{E}[(\underbrace{\mathbf{z}(\mathbf{p})}_{\mathbf{z}(\hat{\mathbf{p}})+\mathbf{J}(\mathbf{p}-\hat{\mathbf{p}})} - \underbrace{\mathrm{E}[\mathbf{z}(\hat{\mathbf{p}})])}_{\mathbf{z}(\hat{\mathbf{p}})}(\mathbf{z}(\mathbf{p})-\mathrm{E}[\mathbf{z}(\mathbf{p})])^T] \\
&\approx \mathrm{E}[\{\mathbf{J}(\mathbf{p}-\hat{\mathbf{p}})\}\{(\mathbf{J}(\mathbf{p}-\hat{\mathbf{p}})\}^T] \\
&\approx \mathbf{J}\{\underbrace{\mathrm{E}[(\mathbf{p}-\hat{\mathbf{p}})(\mathbf{p}-\hat{\mathbf{p}})^T]}_{\mathrm{Cov}(\hat{\mathbf{p}})}\}\mathbf{J}^T \\
&\approx \mathbf{J}\mathbf{C}_{pp}\mathbf{J}^T \equiv \mathbf{C}_{\hat{z}\hat{z}} && (4.11)
\end{aligned}$$

If no correlations are taken into account, i.e., if $\mathbf{C}_{pp}$ is a diagonal matrix with the parameter variances $\sigma^2_{p_j}\quad (j = 1, \ldots, n)$ on its diagonal, the uncertainty of a specific model prediction $\hat{z}_i\quad (i = 1, \ldots, m)$, e.g., the predicted pressure at a certain point in space and time, is given by:

$$\sigma^2_{\hat{z}_i} = \left(\frac{\partial z_i}{\partial p_1}\right)^2 \sigma^2_{p_1} + \left(\frac{\partial z_i}{\partial p_2}\right)^2 \sigma^2_{p_2} + \cdots + \left(\frac{\partial z_i}{\partial p_n}\right)^2 \sigma^2_{p_n} \qquad (4.12)$$

The variance of a model prediction is the sum of the squared products of the partial derivatives times the variance of the respective parameter. In general, the covariance matrix of the predicted system response is the expected value of the squared differences between the prediction and the expected value of the prediction.

4.2.3 Monte Carlo simulations

An alternative to First-Order-Second-Moment error propagation analysis is performing *Monte Carlo* simulations. Monte Carlo (MC) simulation requires repetitive solution of the forward problem, with the parameters randomly sampled from their suspected probability distributions. The output from MC runs is then used to analyze the statistical properties of the distribution of the model prediction. The procedure is as follows:

1. Define probability distributions for all parameters.

2. Randomly sample parameter values from the defined distributions, i.e., generate parameter values that follow the given probability density function.

3. Combine sampled parameter values randomly to obtain a parameter vector. Since the combination of parameter values is random, no correlation between parameters is introduced.

4. Run simulation and store results.

5. Repeat steps (2) through (4) many times.

6. Perform statistical analysis (histogram, moments, etc.) of ensemble of model output.

Advantages:

- Any distribution function (uniform, normal, log-normal, exponential, etc.) can be chosen to describe parameter uncertainty.

- No assumption is made about the distributional form of the model output, i.e., the full distribution of prediction uncertainty is obtained. Monte Carlo is a "full distribution analysis."

- Nonlinearities are automatically taken into account.

- Results from Monte Carlo simulations are physically feasible (note that FOSM analysis assigns a certain probability also to system behaviors that are physically not possible!).

Disadvantages:

- Monte Carlo simulations are computationally very expensive.

- Results from MC uncertainty analysis are difficult to report (note that FOSM assumes that the errors are normally distributed, i.e., they can be reported by a single number, namely the standard deviation).

- Correlations among parameters are not taken into account (note that extensions of the simple Monte Carlo approach exist that take into account correlations among parameters).

4.2.4 Example

A short example illustrates the differences between the linear FOSM uncertainty propagation analysis and Monte Carlo simulations. As discussed above, for small standard deviations of the input parameters, and if the model output can be approximated by a linear function of the parameters within the range of the error band, FOSM is a fast method to calculate a measure of prediction uncertainty that is easy to report. If the model is highly nonlinear, and the uncertainties of the input parameters are large, Monte Carlo simulations have to be performed to examine many parameter combinations according to their probabilities. Monte Carlo simulations provide the full distribution of the model output at the selected points in space and time. The Monte Carlo method is very flexible in handling non-Gaussian distributions of both input parameters and output variables, but they are computationally expensive, and results are difficult to report. In this section we compare both approaches for a synthetic laboratory experiment consisting of three parts: (1) injection of water into a partially saturated sand column for 5 minutes under constant pressure, (2) injection of gas for 2.5 minutes, followed by (3) a 2.5 minute shut-in recovery period.

The standard deviations of three uncorrelated input parameters, the logarithm of the absolute permeability $\log(k)$, porosity ϕ, and the initial gas saturation S_{g_i} of a soil column are assumed to be 0.1, 0.05, and 0.05, respectively. Performing a simulation of a synthetic laboratory experiment, we are interested in the reliability of the model predictions, for example, the uncertainty of the pressure in the center of the column.

The results from both the FOSM and Monte Carlo uncertainty analyses are visualized in Figure 4.5. While the linear FOSM analysis gives a reasonable estimate of prediction uncertainty for most parts of the experiment, the Monte Carlo simulations reveal an asymmetry of the output distribution in the period where nonlinearities prevail. Note that FOSM analysis assigns a certain probability to pressure responses that are below 1 bar, which is physically not possible in this experiment. The Monte Carlo simulations stay away from this lower bound. A parameter combination of low permeability, high porosity, and low initial gas saturation yielded the highest pressures.

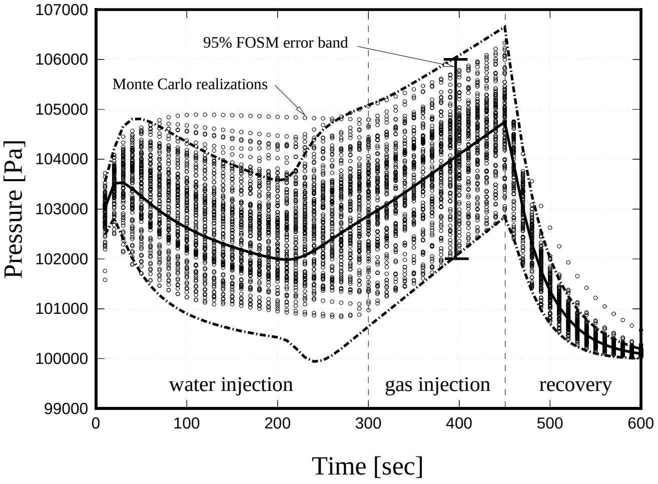

Figure 4.5: Comparison between FOSM and Monte Carlo uncertainty propagation analyses.

Chapter 5

Computer Exercise

5.1 Purpose

The purpose of the following computer exercises is to make the procedure of inverse modeling transparent. Seemingly abstract concepts such as the stochastic model are more easily understood when applying them to actual data. On the other hand, obtaining a good match and correctly interpreting inverse modeling results remains a difficult task especially when applying to real data. One should keep in mind, however, that the same difficulties must be faced when trying to match data by trial-and-error model calibration. The additional information provided by inverse modeling reveals weaknesses in the current model conceptualization, uncovers high sensitivities that need to be carefully examined, and points towards aspects of the model that are supposed to be modified.

The inverse modeling code provided for these exercises is iTOUGH2 [6][7],[1] which is based on the TOUGH2 simulator [17][18].

5.2 iTOUGH2

5.2.1 Summary description

iTOUGH2 is a computer program that provides inverse modeling capabilities for the TOUGH2 code. TOUGH2 is a numerical simulator for non-isothermal flows of multicomponent, multiphase fluids in porous and fractured media. While the main purpose of iTOUGH2 is to estimate model-related hydraulic properties by calibrating TOUGH2 models to laboratory or field data, the information obtained by evaluating parameter sensitivities can be used to optimize the design of an experiment, and to analyze the uncertainty of model predictions.

[1] Information about iTOUGH2 can be obtained from the World Wide Web at `http://www-esd.lbl.gov/iTOUGH2`.

iTOUGH2 solves the inverse problem by automatic model calibration based on the maximum likelihood approach. All TOUGH2 input parameters can be considered unknown or uncertain. The parameters are estimated based on any type of observations for which a corresponding TOUGH2 output is available. A number of different objective functions and minimization algorithms are available. One of the key features of iTOUGH2 is its extensive error analysis which provides statistical information about residuals, estimation uncertainties, and the ability to discriminate among model alternatives. The impact of parameter uncertainties on model predictions can be studied by means of First-Order-Second-Moment uncertainty propagation analysis or Monte Carlo simulations.

5.2.2 Basic elements of iTOUGH2 input language

General Remarks

- provide TOUGH2 input file in standard TOUGH2 format
- provide iTOUGH2 input file defining parameters, data, and program options
- structured high-level command input language
- free format, case-insensitive, flexible command interpreter, comments, error messages

Elements

`>, >>, >>>, ...`	: Command Level Marker: Command expected on same line
`<, <<, <<<, ...`	: Terminate command level
Commands/Keywords	: Trigger option, request input, invoke new command level
Parameters	: Numerical values, strings, variable names, etc.
:	: Provide input parameter(s) immediately after a colon

Example

```
> PARAMETER
  >> estimate ABSOLUTE permeability
     >>> ROCK: ATMOS BOUND SOI_1 +3  (one value for 6 rocks)
         >>>> print LIST of all possible commands
         >>>> ANNOTATION       : permeability
         >>>> estimate LOGARITHM
         >>>> INDEX            : 1 2 (horizontal perm. only)
         >>>> initial GUESS is : -14.0
         >>>> RANGE : -18.0  -12.0   (upper and lower bounds)
         >>>> maximum STEP     : 1.0 (please HELP!)
         >>>> don't WEIGHT p.i.: 0.0
         <<<<
     <<<
  <<
```

Special Commands

The following special commands are applicable on all command levels.

`>>> LIST`	: prints list of all commands accepted on this command level
`>>> command HELP`	: provides short help message about the command
`/*`	: beginning of ignored block
`*/`	: end of ignored block
`#` in first column	: line ignored

Example Special Commands

```
  >> LIST all available commands on this level
  >> PRESSURE (which pressure is this? --> HELP!!!)
     >>> ELEMENT: AX1_1
#        >>>> ANNOTATION: L1 SOURCE 1
         >>>> DATA [MINUTE]  FILE: hL1.1.dat
         >>>> DEVIATION: 100.0 HELP
         <<<<
/*            ignore the following 5 lines
     >>> ELEMENT: AX112
         >>>> ANNOTATION: L4 DETECT
         >>>> DATA [MINUTE] FILE: hL4.1.dat
         >>>> DEVIATION: 100.0
         <<<<
*/
     <<<
  <<
```

First Level Commands and General Structure

```
> PARAMETERS  (Define TOUGH2 parameters to be estimated)
  >> specify parameter type
     >>> specify parameter domain
         >>>> provide details
         <<<<
     <<<
  <<


> OBSERVATION (= data based on which parameters will be estimated)
  >> specify calibration points in TIME
  >> specify observation type
```

```
     >>> specify location
         >>>> provide details
         >>>> provide data
         <<<<
     <<<
  <<

> COMPUTATION (program options)
  >> Stopping criteria
  >> Program options
  >> Output
```

All iTOUGH2 commands are documented on the iTOUGH2 Web Site `http://www-esd.lbl.gov/iTOUGH2`, or can be printed to the screen using command `it2help`.

5.3 Sample Problem

5.3.1 Problem description

Consider the following laboratory experiment (Figure 5.1):

Water is injected applying a constant pressure head of 1 m into a one-dimensional, horizontal column filled with uniform, partially saturated sand. Pressure at the outlet is kept constant at atmospheric conditions.

We assume that the objective of the laboratory experiment is to estimate the permeability and the porosity of the sand as well as the initial gas saturation. Furthermore, we presume that only one flow meter and one pressure transducer are available for data collection. The injection rate is measured at the inlet, and pressure measurements are taken at the center of the column. The measurement uncertainties of the two instruments are 5 ml/min and 200 Pa, respectively.

Three files holding synthetically generated sets of flow rate and pressure data with no or random measurement errors are provided on files *nonoise.dat*, *noisy.dat*, and *noisier.dat*, respectively. The experiment can be simulated using TOUGH2 input file *darcy*. TOUGH2 is used in combination with the Equation-Of-State (EOS) module number 3 for water, air, and heat.

5.3.2 Exercise 1: Solve the forward problem

In a first step, we solve the forward problem with some initial guesses for the unknown parameters. Type:

Figure 5.1: Schematic of a transient, two-phase flow experiment. Water is injected into a partially saturated column. Injection rate at the inlet and pressure in the center of the column are measured as a function of time.

```
    itough2 darcy1i darcy 3 &
```

The iTOUGH2 output file *darcy1i.out* or the plot file *darcy1i.col* contain information about the differences between the data and the simulation with the initial parameter set. Answer the following questions:

Questions

1. Generate a plot of the data and the simulation results (see also Figures 5.2 and 5.3 on page 43), look at the residual plot in file *darcy1i.out*, or examine the columns under heading `RESIDUAL ANALYSIS`. Describe the mismatch between the data and the simulation.

2. How would you change the parameters to improve the match?

5.3.3 Exercise 2: Calibration, no measurement errors

Run the following inversion:

```
itough2 darcy2i darcy 3 &
```

Type `prista` during the inversion to check the status of your run. Type

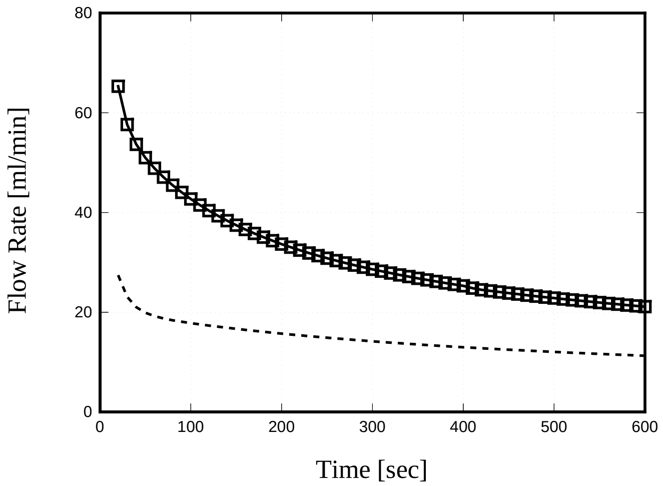

Figure 5.2: Flow rates at inlet calculated with initial parameter set (dashed line) and after calibration (solid line). Synthetic data are shown as squares.

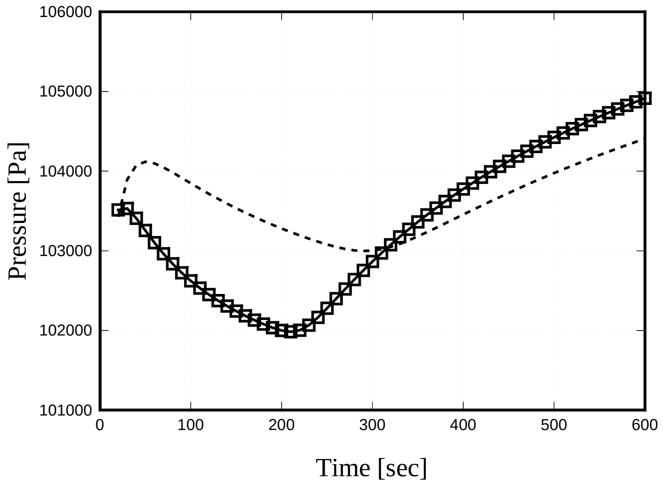

Figure 5.3: Pressure transient at center of column calculated with initial parameter set (dashed line) and after calibration (solid line). Synthetic data are shown as squares.

`kit` for additional options and early termination of your run. Examine file *darcy2i.out* and comment on the following questions:

Questions

1. Which parameters have been estimated?
2. How many calibration points were selected?
3. Define and indicate the degree of freedom of this inversion.
4. Which minimization algorithm was used?
5. Describe the path taken by the minimzation algorithm.
6. Which parameter is the most sensitive one?
7. Which single observation (type and time) contains the most information regarding each of the parameters, and overall?
8. Which observation type (flow rates or pressures) contains the most information regarding each of the parameters, and overall?
9. What is the value of the *a posteriori* error variance s_0^2? Why is it so small?
10. Why was the error analysis based on the *a priori* error variance σ_0^2?
11. What is the estimation uncertainty of the three parameters?
12. What does the correlation coefficient between $\log(k)$ and porosity indicate?
13. Give a physical explanation why the correlation coefficient is positive.
14. Which of the three parameters is the most independent, and which exhibits the largest overall correlation?
15. Examine and discuss the correlation chart.
16. What is the inital and final value of the objective function?
17. What is the result of the Fisher model test?
18. Why is the final value of the objective function so small?
19. What is the best estimate parameter set and the estimation uncertainty?

5.3.4 Exercise 3: Calibration, noisy data

Copy iTOUGH2 input file *darcy2i* to *darcy3i.* Edit file *darcy3i.* Change name of the selected data from data file *noisy.dat* or *noisier.dat* instead of *nonoise.dat* by reassigning the comment character "`#`". Rerun the inversion.

Comment on the following issues:

1. How big is the *a posteriori* error variance?

2. Elaborate on the result of the Fisher Model test.

3. Which observation leads to the maximum residual? Is it acceptable?

4. What is the best estimate parameter set? Compare it with the true parameter set obtained in the previous inversion.

5. Comment on the fact that porosity is overestimated, whereas initial gas saturation in underestimated.

6. Under which conditions do you expect predictions made with the estimated parameter set to be acceptable?

5.3.5 Exercise 4: Explore

File *darcy4i* (see following page) is a template based on file *darcy3i.* Replace the question marks in this file, and explore the effects of changing options on the results. You may introduce fewer or more parameters than used before, omit or add data, change standard deviations of the data or the initial parameter guesses, use different objective functions or minimization algorithms, and change iteration parameters. Try to explain the changes with respect to the reference cases discussed in Sections 5.3.3 and 5.3.4. Use the command index provided on the iTOUGH2 Web page `http://www-esd.lbl.gov/iTOUGH2`.

```
--------------------------------------------------------------------------------
                           iTOUGH2 SAMPLE PROBLEM
--------------------------------------------------------------------------------
Direct problem      : darcy
Inverse problem     : darcy4i
Data files          : nonoise.dat, noisy.dat, noisier.dat
EOS module          : 3
Description         : Estimate permeability, porosity, and initial gas saturation
                      based on synthetic pressure and flow rate data.
Execution           : itough2 darcy4i darcy 3 &
--------------------------------------------------------------------------------

Look for "???" and fill in parameters as needed.
Invoke, modify, or remove options using "#" in the first column,
adding or removing the comment character /*,
or by adding or removing the appropriate command level marker >>, >>>, ... .
Check out additional options listed on the Web at http://www-esd.lbl.gov/iTOUGH2,
click on "Command Index" and the individual commands for a description. */


> PARAMETER

/* ??? (delete this line if you want to estimate absolute permeability)
  >> ABSOLUTE permeability
     >>> MATERIAL             : SAND_ BOUND
         >>>> ANNOTATION      : log(abs. perm.)
         >>>> estimate LOGARITHM
         >>>> initial GUESS   : -??.?
         >>>> RANGE           : -13.0 -11.0
         >>>> VARIATION       :   0.5
         <<<<
      <<<
*/

/* ??? (delete this line if you want to estimate porosity)
  >> POROSITY
     >>> MATERIAL             : SAND_
         >>>> ANNOTATION      : Porosity
         >>>> estimate VALUE
         >>>> RANGE           : 0.01 0.99
         >>>> initial GUESS   : 0.??
         >>>> VARIATION       : 0.05
         <<<<
     <<<
*/

/* ??? (delete this line if you want to estimate initial gas saturation)
  >> INITIAL condition for primary variable No.:2
     >>> DEFAULT
         >>>> ANNOTATION      : Initial gas sat.
         >>>> VALUE
         >>>> admissible RANGE : 10.01 10.99
         >>>> initial GUESS   : 10.?? (you must add 10.0 to saturation!)
         >>>> VARIATION       :  0.05
         <<<<
     <<<
*/

  <<
```

Block > PARAMETER of iTOUGH2 input file *darcy4i*.

```
> OBSERVATION

... You may change the number of calibration points and
    their spacing in time.

  >> select : 20 points in TIME, EQUALLY spaced between
     30.0 600.0 seconds

/* ??? (delete this line if you want to match pressure data)
  >> PRESSURE
     >>> ELEMENT                  : A1125
         >>>> ANNOTATION          : Pressure 1/2
         >>>> HEADER contains     : 5 lines
         >>>> SET                 : 1
         >>>> COLUMNS             : 1 2

... You may select either a set with noisy data (noisy.dat),
    a noisier data set (noisier.dat), or one
    with exact measurements (nonoise.dat)
         >>>> Read DATA from FILE : noisy.dat ???  (time is in MINUTES)

         >>>> standard DEVIATION  : ??? Pa         (measurement error)
         <<<<
     <<<
*/

/* ??? (delete this line if you want to match flow rate data)
  >> LIQUID FLOW RATE
     >>> CONNECTION defining inlet: IN__0  A11_1
         >>>> ANNOTATION          : Flow inlet
         >>>> FACTOR              :-1.666667E-05 (ml/min - kg/sec)
         >>>> HEADER contains     : 5 lines
         >>>> SET                 : 2
         >>>> COLUMNS             : 1 2
         >>>> Read DATA from FILE : noisy.dat  ??? (time is in MINUTES)
         >>>> standard DEVIATION  : ??? ml/min     (measurement error)
         <<<<
     <<<
*/
  <<
```

Block `> OBSERVATION` of iTOUGH2 input file *darcy4i*.

```
> COMPUTATION

  >> STOPPING criteria
     >>> after :  5 ITERATIONS
     >>> initial value of LEVENBERG parameter: 0.001
     >>> MARQUARDT parameter: 10.0
     >>> maximum scaled STEP size: 2.0
         ignore WARNINGS
     <<<

  >> OUTPUT
         RESIDUALS
     >>> FORMAT of plot file *.col: COLUMNS
     >>> output time in MINUTES
     <<<

  >> JACOBIAN
         CENTERED finite differences
     >>> FORWARD finite differences
     >>> PERTURB by:  1 %
     <<<

  >> ERROR
         use A POSTERIORI error variance
         use A PRIORI error variance
     >>> perform FISHER model test
     >>> significance level (1-ALPHA) = 95 %
         perform FOSM uncertainty propagation analysis
     <<<

  >> OPTIONS

--- You may choose a different objective function:
     >>> LEAST-SQUARES
         L1-ESTIMATOR
         QUADRATIC-LINEAR, k : 2
         CAUCHY
         ANDREW c : 1.5

--- You may choose a different minimization algorithm:
         SIMPLEX
     >>> LEVENBERG-MARQUARDT
         GAUSS-NEWTON
     <<<
  <<
<
--------------------------------------------------------------------------------
```

Block `> COMPUTATION` of iTOUGH2 input file *darcy4i*.

5.3.6 Additional exercises

Additional exercises (Sample Problems 1 through 7) are part of the iTOUGH2 distribution and can be found in subdirectories `~/itough2/samples/sampleX`.

1. Sample Problem 1 is a tutorial similar to the one discussed in this lecture. It covers the main applications:

   - Solving the forward problem;
   - Performing a sensitivity analysis for optimizing the experimental design;
   - Estimating parameters by automatic model calibration;
   - Assessing prediction uncertainties by means of linear uncertainty propagation analysis and Monte Carlo simulations.

2. Sample Problem 2 discusses the analysis of the gas-pressure-pulse-decay experiment described in Sections 1.3, 4.1.3, and 4.1.4 as well as in [9]. Parameter correlations, the problem of non-uniqueness, and the parameterization of systematic errors are discussed in detail.

3. Sample Problem 3 demonstrates the use of sensitivity measures for automatically selecting the parameters to be estimated based on synthetically generated data from a fractured geothermal reservoir.

4. Sample Problem 4 features different minimization algorithms for the analysis of data from a multi-step, radial desaturation experiment.

5. Sample Problem 5 shows the matching of saturation, water potential, and pneumatic data from deep boreholes, where the observed gas pressure fluctuations are a result of atmospheric pressure variations.

6. Sample Problem 6 discusses the analysis of a ventilation experiment described in [8].

7. Sample Problem 7 examines numerical diffusion effect and illustrates the estimation of *model-related* parameters.

All sample problems are described in [7].

Bibliography

[1] Beck, J. V., and K. J. Arnold, *Parameter Estimation in Engineering and Science*, John Wiley & Sons, New York, 1977.

[2] Carrera, J., and S. P. Neuman, Estimation of aquifer parameters under transient and steady state conditions, 1, Maximum likelihood method incorporating prior information, *Water Resour. Res., 22*(2), 199–210, 1986.

[3] Carrera, J., and S. P. Neuman, Estimation of aquifer parameters under transient and steady state conditions, 2, Uniqueness, stability, and solution algorithms, *Water Resour. Res., 22*(2), 211–227, 1986.

[4] Carrera, J., and S. P. Neuman, Estimation of aquifer parameters under transient and steady state conditions, 3, Application to synthetic and field data, *Water Resour. Res., 22*(2), 228–242, 1986.

[5] Finsterle, S., iTOUGH2 user's guide *Lawrence Berkeley Natl. Lab. Rep., LBNL-40040*, Berkeley, CA, 1999.

[6] Finsterle, S., iTOUGH2 command reference, *Lawrence Berkeley Natl. Lab. Rep., LBNL-40041 (updated)*, Berkeley, CA, 1999.

[7] Finsterle, S., iTOUGH2 sample problems, *Lawrence Berkeley Natl. Lab. Rep., LBNL-40042 (updated)*, Berkeley, CA, 1999.

[8] Finsterle, S. and K. Pruess, Solving the estimation-identification problem in multiphase flow two-phase flow modeling, *Water Resour. Res., 31*(4), 913–924, 1995.

[9] Finsterle, S. and P. Persoff, Determining permeability of tight rock samples using inverse modeling, *Water Resour. Res., 33*(8), 1803–1811, 1997.

[10] Finsterle, S. and J. Najita, Robust estimation of hydrogeologic model parameters, *Water Resour. Res., 34*(11), 2939–2947, 1998.

[11] Gill, P. E., W. Murray, and M. H. Wright, *Practical Optimization*, Academic, San Diego, Calif., 1981.

[12] Jacoby, S. L. S., J. S. Kowalik, and J. T. Pizzo, *Iterative Methods for Nonlinear Optimization Problems*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1972.

[13] Klinkenberg, L. J., The permeability of porous media to liquids and gases, in *Drilling Production Practice*, pp. 200–213, Am. Pet. Inst., Washington, D. C., 1941.

[14] Kool, J. B., J. C. Parker, and M. Th. van Genuchten, Parameter estimation for flow and transport models—a review, *J. Hydrol.,91*, 255–293, 1987.

[15] Pike, R. W., *Optimization for Engineering Systems*, Van Nostrand Reinhold Company Inc., New York, 1986.

[16] Marquardt, D. W., An algotithm for least squares estimation of nonlinear parameters, *SIAM J. Appl. Math., 11*, 432–441, 1963.

[17] Pruess, K., TOUGH user's guide, *U.S. Nucl. Regul. Comm. Rep., NUREG/CR-4645*, 1987.

[18] Pruess, K., TOUGH2—A general-purpose numerical simulator for multiphase fluid and heat flow, *Lawrence Berkeley Natl. Lab. Rep., LBL-29400*, Berkeley, CA, 1991.

[19] Yeh, W. W.-G., Review of partameter estimation procedures in groundwater hydrology: The inverse problem, *Water Resour. Res., 22*(2), 95–108, 1986.